中文文献题录数据机构名称归一化研究

■ 杨昭[1] 任娟[2,3]
[1] 上海交通大学图书馆 上海 200240 [2] 上海出版印刷高等专科学校 上海 200093
[3] 上海出版传媒研究院 上海 200093

摘 要：[目的/意义]大数据时代，机构名称数据呈现海量性、动态性、多样性等新特征，机构名称归一化可改善大数据环境下科研管理、学科评价、学科服务中的数据可靠性，提升基于机构名称的数据检索质量和应用效果。[方法/过程]从语言学角度和模型构建层面研究机构名称归一化，构建基于共现关系和相似度的机构名称归一化框架模型，提出机构名称实体边界识别方法，编制机构多层级词表，提出机构名称归一化方法，最后选取2008－2018年中文文献题录数据进行实验。[结果/结论]实验结果验证了模型的有效性，对其他类型机构名称归一化有一定的启发。

关键词：机构名称 归一化 模型构建 大数据 实体边界识别
分类号：G254
DOI：10.13266/j.issn.0252－3116.2020.04.011

1 引言

机构名称是机构基本属性、内在规律以及特殊性的综合反映。机构名称包括规范名称、曾用名、译名、合并名称、附属独立名称等，可区分为规范名称和变异名称两类。其中，规范名称是指依据国家标准规范等由权威机构发布的某一机构实体的名称；变异名称是指同一机构实体的多种名称表达，主要有全称简称、中文简繁体名称、多语言形式译名、著录错误名称、不同数据源和不同时间段的名称等。机构名称归一化旨在将同一机构实体名称的不同表达形式集中起来，建立规范名称与变异名称之间的对应关系，通过赋予机构唯一标识符的方式达到机构识别的目的[1]。实质上，归一化就是识别机构实体间的同一关系、相继关系、隶属关系等。大数据时代，随着学术大数据涌现，机构更迭频繁，文献数据著录不规范，机构名称数据呈现海量性、动态性、多样性等新特征，机构名称归一化可改善大数据环境下科研管理、学科评价、学科服务中数据可靠性，提升基于机构名称的数据检索质量和应用效果。机构名称归一化是建立机构名称规范档、规范库的核心和关键，是机构知识库建设的重要内容，是图书馆开展学科服务的基础和前提，是提高检索查询的查全率和查准率的重要手段。

就主要文献数据库的机构扩展检索功能而言，多数文献数据库都具有“作者＋作者单位”的筛选功能，但没有上下级隶属关系的区分[2]。WOS数据库和SCOPUS数据库都提供机构扩展检索功能，可实现一级机构名称归一化；维普数据库、CNKI数据库和万方数据库可提供期刊论文的作者机构字段，其中维普数据库标注作者与机构的对应关系，CNKI数据库未标注作者与机构的对应关系，万方数据库利用XML文档标注作者与一级机构的对应关系。以上数据库都未涉及二级以下机构名称归一化。

就机构名称归一化的研究方法而言，学者们分别提出了基于规则的方法和基于统计的方法。前者主要是利用机构名称结束标识词触发的形式识别命名实体边界，总结组合规则、语义模式和语法特征，通过关键词进行识别。由于缺乏各类型机构名称之间的语义关系定义，仅仅依赖于名称形式上匹配，会出现漏统计、错误统计等问题[3]。后者主要是在语料库的基础上，采用文本特征结合机器学习算法的范式进行识别。然而，机器学习算法都是黑箱的，模型对于结果缺乏可解释性，出错时难以发现错误原因和提出修正策略[4]。

就中英文机构名称著录差异而言，中文机构名称一般不包含英文机构名称的空格分隔符和英文多层级机构名称间的逗号分隔符等，需要进行自动分词和命

作者简介：杨昭（ORCID：0000－0003－1803－3516），馆员，硕士，E-mail：zhaoyang2017@sjtu.edu.cn；任娟（ORCID：0000－0002－1814－9378），副主任，副教授，博士。
收稿日期：2019－06－21 修回日期：2019－10－15 本文起止页码：95－102 本文责任编辑：杜杏叶

名实体边界识别。

本文从语言学角度和模型构建层面研究机构名称归一化，首先对机构名称进行语法、语义特征分析，进而提出机构名称组合规则；构建机构名称归一化模型；提出机构名称的实体边界识别方法；编制机构多层级词表，提出机构名称归一化方法；最后实证检验模型的有效性。

2　相关研究

国内外学者针对机构名称规范化问题，从机构识别角度提出了多种方法和策略，主要有以下几个方面：一是基于规则的方法。沈嘉懿等人针对网络文本数据提出了基于规则的中文机构识别方法，通过机构后缀词库、规则匹配和贝叶斯模型识别右边界和左边界[5]。杨波等人针对 WOS 题录数据提出了基于规则和统计的机构名称映射算法[6]。二是基于统计的方法。胡万亭等人利用百度百科词条，提出了基于词频统计的机构名称识别方法[7]。买合木提·买买提等人提出了一种基于条件随机场模型的维吾尔文机构名称识别方法[8]。杨瑞仙和毛一雷提出了一种基于规则与向量空间模型相结合的科研机构命名识别方法[9]。三是中文机构名称归一化。贾君枝等人利用 CNKI 数据库构建了机构名称特征词表，提出了基于 TF-IDF 和 K 均值聚类算法的中文机构名称归一化方法[1]。杨奕虹等人在万方数据库的基础上，采用叙词表的知识组织方式，构建了中文机构多层级词表[10]。曾建勋和贾君枝引入 Schema 词汇表构建了机构名称规范数据的语义模型[3]。孙海霞等人利用中文生物医学文献数据库作为语料库，提出了基于 K 均值聚类算法的中文机构名称归一化方法[11]。

3　机构名称归一化框架模型与算法介绍

区别于基于网络文本的机构命名实体识别，针对中文文献题录数据的机构名称归一化，其难点在于缺乏关联数据条件下识别机构实体间的同一关系、相继关系和隶属关系等。就机构名称归一化的实现路径而言，利用人工编制的机构多层级词表进行机构映射具有准确性优势，利用字符串相似度可自动聚类因著录错误等产生的变异名称，利用机构名称的层级结构可识别上下级隶属关系。本文针对机构名称数据呈现的海量性、动态性、多样性等新特征，采用机构映射的机构归一化策略，将机构识别转换为精确匹配问题，提出一种基于共现关系和相似度的机构名称归一化框架模型（见图 1）。该框架模型的基本假设：变异名称与规范名称之间仅存在一对一或多对一的映射关系；其基本思想：利用机构名称的层级结构和层级共现关系实现隶属关系和同一关系的联合式机构识别，并迭代处理整个过程。首先，将分词和命名实体边界识别结合在一起，提出机构名称实体边界识别方法，识别隶属关系；之后，基于规则、共现矩阵和相似度构建中文机构多层级词表，识别同一关系和隶属关系；最后，提出一种基于精确匹配的机构名称归一化方法，实现模型泛化。

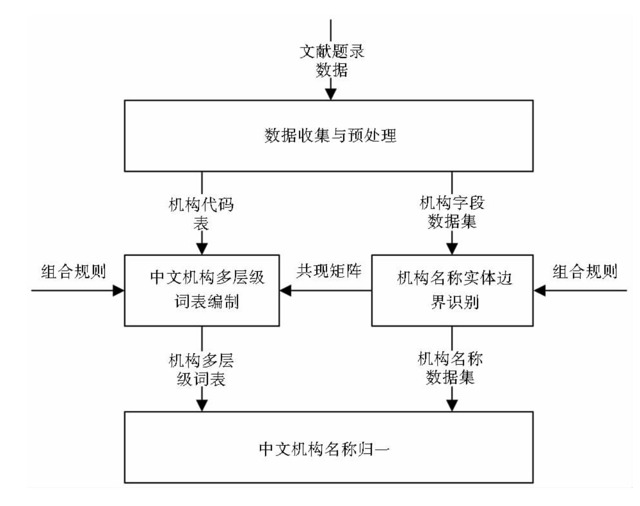


图 1　机构名称归一化框架模型

3.1　机构名称数据收集与预处理

数据收集与预处理包括选择数据源、特殊标点符号预处理和提取机构字段 3 个步骤：①选择数据源。中文机构名称数据来源主要为维普数据库、CNKI 数据库和万方数据库。选择维普数据库作为数据来源。②特殊标点符号预处理。由于机构著录不规范，存在以空格和“、”“/”“（”“）”等特殊字符作为机构边界标识符，预处理时需要将其替换“;”。③提取机构字段。在维普数据的机构字段中，左边界分隔符一般为“]”，而右边界分隔符为“;”，中间分隔符为“,”，字段包含一个或多个层级的机构全称、国别、城市、邮编、地址等。经过机构字段预处理，得到机构字段数据集，见表 1。

3.2　机构名称实体边界识别

机构名称的一般表达式为：$F + M^{*} + S$。利用 NLPIR 工具进行分词标注和词频统计，可获得机构结束标识词。以高校为例，机构名称组合规则及结束标识词，见表 2。

机构层级组合规则的一般表达式为：一级机构 +［二级机构］+［三级机构］+ … +［N 级机构］。基于机构字段数据集，利用形如“select count（ID）from table where organization like‘%大学%’and organization

表 1　中文文献题录数据机构字段的预处理结果(样例)

ID	机构序号	作者	机构名称	城市	邮编	国别
VIP:673485847	1	张瑶[1];丁桂甫[1];王强[2];张丛春[1];程萍[1]	上海交通大学电子信息与电气工程学院微纳电子系微纳米加工技术重点实验室	上海	200240	中国
VIP:673485847	2	张瑶[1];丁桂甫[1];王强[2];张丛春[1];程萍[1]	上海中航商用航空发动机制造有限责任公司	上海	201108	中国

表 2　高校机构名称组合规则

机构等级	首词 F	中间词 M	后缀词 S
一级	[国名|地名|人名|方位词|主管部门|专造名]	[学科|行业|活动内容|序数词]*	[大学|学院|学校]
二级	[国名|地名|人名|方位词|主管部门|专造名]*	[学科|行业|活动内容|序数词]*	[分校|学院|学校|中学|小学|系|中心|研究院|研究所|图书馆|档案馆|博物馆|部|处|办公室|委|会|有限公司|医院]
三级	[国名|地名|人名|方位词|主管部门|专造名]*	[学科|行业|活动内容|序数词]*	[学院|系|研究院|研究所|实验室|中心|部|处|科|室|党委|办公室|编辑室|研究室|监督室|技术室|设计室|美术馆|体育馆|党校|委员会|基地|科技园|有限公司|国资办|选培办|办事处|幼儿园|医院|队|站|所|厂]
四级	[国名|地名|人名|方位词|主管部门|专造名]*	[学科|行业|活动内容|序数词]*	[系|实验室|研究所|中心|办公室|部|处|学校|编辑室|医院]

like‘%学院%’ and organization like‘%系%’;”的 SQL 查询，统计机构结束标识词的二重和三重共现频次，生成层级组合规则。以高校为例，机构层级组合规则及结束标识词为：①大学＋[学院|部|处]＋[系|办公室]＋[实验室|研究院|设计院|研究所|中心|基地]；②大学＋[学院|部|处]＋[系|办公室]；③大学＋[学院|部|处]；④大学＋[系|办公室]；⑤大学＋[学院|部|处]＋[实验室|研究院|设计院|研究所|中心|基地]；⑥大学＋[系|办公室]＋[实验室|研究院|设计院|研究所|中心|基地]；⑦大学＋[实验室|研究院|设计院|研究所|中心|基地]；⑧[学院|学校]＋[系|办公室]＋[实验室|研究院|设计院|研究所|中心|基地]；⑨[学院|学校]＋[系|办公室]；⑩[学院|学校]＋[实验室|研究院|设计院|研究所|中心|基地]；⑪学院＋[学院|实验室|研究院]。

本文将分词和命名实体边界识别结合在一起，提出一种机构名称实体边界识别算法。输入：机构字段数据集和机构层级组合规则。输出：机构名称数据集。流程如下：

步骤 1：一级机构识别

采用基于等值匹配的分块算法[12]，将实体属性中的机构名称、国别、地址、城市、邮编定义为五个分块键，构造数据记录过滤条件，对机构字段数据集进行分组，实现一级机构的识别。

步骤 2：词性标注

利用 NLPIR 工具，选择中科院二级标注集作为词性标注集。经词性标注后的结果，如“上海/ns 交通/n 大学/n 电子/n 信息/n 与/cc 电气/n 工程/n 学院/n 自动化/vd 系/v 系统/n 控制/vn 与/cc 信息/n 处理/v 教育部/nt 重点/n 实验室/n”。

步骤 3：确定机构后缀词

依据词性标注结果，查找机构字段中的所有机构后缀词。词性标注结果“系/v”和“系统/n”，判定“系/v”为机构原子后缀词；“系统/n”为定语修饰词。最终找到“大学”“学院”“系”“实验室”4 个机构后缀词和“系统”1 个定语修饰词。

步骤 4：确定机构全称右边界

匹配机构层级组合规则，并以“#”作为分隔符标识各级机构名称右边界。上例经右边界标识后的结果为：“上海交通大学#电子信息与电气工程学院#自动化系#系统控制与信息处理教育部重点实验室#”。

3.3　中文机构多层级词表编制

中文机构多层级词表编制的基本步骤如下：

(1)人工收集规范名称，制作机构规范名称基础词表。依据一级机构主页的院系设置、机构设置、历史沿革等栏目，人工整理一级机构和二级以下机构的规范名称，制作机构规范名称基础词表。也可利用百度百科、机构成立新闻报道、机构代码表等提供的机构信息。

(2)识别隶属关系，生成待归并的中文机构多层级词表。提出基于共现关系与等值匹配的分块算法。依据机构层级组合规则，计算二重、三重共现频次，通过设置共现频次阈值，提取机构实体间的共现关系，确定一级机构－二级机构－三级机构的隶属关系。生成未识别同一关系的中文机构多层级词表。

(3)识别同一关系，生成未编码的中文机构多层级词表。利用基于编辑距离的相似度算法，识别同一

关系，归并文本相似的名称，生成未编码的中文机构多层级词表。可将基础词表作为种子加入聚类，提高聚类的效率和质量。

（4）识别相继关系。对于一级机构，进行人工整理，确定其相继关系。主要依据是教育部发布的更名文件和一级机构主页、机构成立新闻报道、百科词条等。对于二级机构，在隶属关系和同一关系识别的基础上，考虑期刊发表周期等，以最终出现年份差距大于 2 作为时序划分标准，以两个二级机构隶属的三级机构个数大于 3 或重合度超过 60% 作为相似性标准，按照以上两个标准确定其相继关系。主要依据是基于以下观察：若一个学院更名，而其下属的所有系、所、实验室并非都更名，依据不更名的系、所、实验室，则可识别出文本不相似的且具有相继关系的两个机构。由于三级机构在文献题录数据中出现频次较低，且需结合作者信息等加以判断，本文将三级机构的相继关系作为同一关系进行处理。

（5）赋予机构的唯一标识符，生成中文机构多层级词表。采用编码的方式，编制基于唯一标识符的中文机构多层级词表。

中文机构多层级词表编制具有以下意义：①词表是文献题录数据与机构名称规范数据之间的映射，词表生成是基于共现矩阵进行去重、合并操作的数据清洗过程。②通过中文机构多层级词表，建立各种变异名称与规范名称的映射关系，实现机构名称的规范化。③采用词表自动编制方法，既保证机构识别的准确度，又节省人力，为机构名称规范建设提供了新的方法。④应用中文机构多层级词表，通过精确匹配，识别海量数据中的机构名称，实现机构名称的归一化。

3.3.1 基于共现关系与等值匹配的分块算法

本文借鉴关联数据的思路，提出一种基于共现关系与等值匹配的分块算法。输入：机构名称数据集。输出：机构名称数据分块结果。流程如下：

步骤 1：建立共现矩阵

在机构名称实体边界识别的基础上，建立各层级机构的二维共现矩阵（大学——学院）和三维共现矩阵（大学——学院——系、所、实验室）。

步骤 2：设置共现频次阈值

通过设置共现频次阈值，提取机构实体间的共现关系。利用共现矩阵同时提取机构实体的隶属关系和共现关系，通过隶属关系揭示机构实体间的语义关系，解决传统的基于规则的方法缺乏各类型机构名称之间的语义关系定义的问题，弥补仅仅依赖于名称形式上匹配的方法缺陷。

步骤 3：分块键定义

采用各层级机构（大学、学院、系）作为实体属性，定义一个或多个分块键，基于机构名称数据在分块键上的键值，将其对应到不同的数据块，提高匹配效率。

3.3.2 基于编辑距离的相似度算法

在数据分块基础上，将作者机构字段排序，采用滑动窗口方法，引入字符串编辑距离算法测度机构名称相似度，滑动窗口宽度设为 30，步长设为 1。机构名称相似度测度采用 Jaro-Winkler 算法计算。该算法是计算两个短字符串之间相似度的一种距离测度算法。

Jaro 距离算法[13]为：

$$\text{Jaro}(str_1, str_2) = \frac{1}{3}\left(\frac{c}{|str_1|}+\frac{c}{|str_2|}+\frac{c-t/2}{c}\right)$$

公式（1）

其中，$|str_1|$ 和 $|str_2|$ 为字符串长度；c 为两个字符串的公共字符数，公共字符需满足 $str_1[i]=str_2[j]$ 和 $|i-j|\leq\frac{1}{2}\min\{|str_1|,|str_2|\}$；$t$ 为变换数，比较两个字符串的第 i 个公共字符，不匹配则为一次变换。

Jaro-Winkler 距离算法[14]为：

$$d_w = d_j + [lp(1-d_j)]$$ 公式（2）

其中，d_j 为两个字符串的 Jaro 距离；l 为前缀相同字符个数，规定其最大值为 4；p 是常数，规定最大为 0.25，Winkler 将其设为 0.1。

3.3.3 基于唯一标识符的多层级词表编制

一级机构唯一标识符包括国家组织机构统一社会信用代码、全国普通高等学校名单中的机构代码等；二级以下机构唯一标识符包括内部机构代码、数据库机构字段编码等。另外，一级机构的电子邮件、二级以下机构的电子邮件、邮编、地址、机构 URL 等具有唯一性且与机构实体具有映射关系的字符串可看作机构唯一标识符。本文以全国普通高等学校名单中的机构代码作为一级机构代码，以五位数字串编码二级以下机构代码，以补充编码作为高校与校外机构共建的且不隶属于其他校内二级机构的协同创新中心、研究院、研究所、研究中心、基地、实验室等的机构代码，并以“UC：”“SC：”“OC：”分别表示一级机构代码、二级以下机构代码、补充编码。

中文机构多层级词表示例见表 3。从表 3 可见，南洋公学和上海交通大学是相继关系；微米/纳米加工技术国防科技重点实验室和电子信息与电气工程学院是隶属关系；微纳电子系和微纳电子学系是同一关系。

杨昭，任娟. 中文文献题录数据机构名称归一化研究[J]. 图书情报工作,2020,64(4):95－102.

表3　中文机构多层级词表示例

序号	变异名称	机构代码	规范名称	父代码	机构级别	成立时间	撤销时间	关系类型
1	南洋公学	UC:10248	上海交通大学		1	1896	1904	相继关系
2	电子信息和电气工程学院	SC:03000	电子信息与电气工程学院	UC:10248	2	2001		同一关系
3	微纳米加工技术重点实验室	SC:34100	微米/纳米加工技术国防科技重点实验室	SC:03000	3	1996		同一关系
4	微米/纳米加工国家级重点实验室	SC:34100	微米/纳米加工技术国防科技重点实验室	SC:03000	3	1996		同一关系
5	微纳电子系	SC:03900	微纳电子学系	SC:03000	3	2014		同一关系
6	薄膜与微细加工技术教育部重点实验室	SC:34200	薄膜与微细技术教育部重点实验室	SC:03900	4	1993		同一关系
7	生物芯片上海国家工程研究中心	OC:00001	生物芯片上海国家工程研究中心	UC:10248	2	2003		同一关系

3.4　机构名称归一

为了实现模型泛化,提出一种基于精确匹配的机构名称归一化算法。输入:机构名称数据集、机构多层级词表。输出:机构名称归一化结果。流程如下:

步骤1:加载中文机构多层级词表,当实体边界识别后的机构数据与中文机构多层级词表精确匹配时,标注其机构代码。

步骤2:当不能精确匹配时,对实体边界识别后的机构数据和中文机构多层级词表的机构名称都进行分词标注,去掉所有标点符号后,再次进行精确匹配,成功匹配后标注其机构代码。

当不能精确匹配时,如机构名称“船舶、海洋与建筑工程学院”的分词标注结果为“船舶/n 、/wn 海洋/n 与/cc 建筑/vn 工程/n 学院/n”,去掉符号后,变异名称转化为“船舶海洋与建筑工程学院”,与规范名称精确匹配成功。

4　实验及结果分析

4.1　数据收集与预处理

为验证模型的有效性,选用维普数据库作为数据源,以上海交通大学作为大学命名机构进行检索,采用检索式检索方式,检索式为“S=(上海交通大学 OR 上海交大 OR 15 个附属医院)”,时间跨度 2008－2018 年。共检索出的文献数量为 145 538 篇,检索日期为 2019 年 3 月 21 日。经数据预处理后获得 233 998 条机构名称数据,以上海交通大学作为第一机构发表期刊论文共计 121 065 篇。

为了评估模型在不同类型机构的适用性,以大学命名和学院命名两种类型的样本检验研究方法的有效性,进一步采用以常熟理工学院作为学院命名机构进行检索,共检索出其 2008－2018 年的文献数量为8 292 篇。

4.2　机构名称实体边界识别

利用 NLPIR 工具进行分词标注,依据机构层级组合规则,对机构名称进行实体边界识别。上海交通大学名称实体边界识别部分结果,如表4所示:

表4　上海交通大学名称实体边界识别部分结果

ID	机构序号	机构名称
VIP:34474797	3	#上海交通大学#第六人民医院#肾内科#
VIP:34474797	10	#上海交通大学#医学院#新华医院#肾内科#
VIP:34474797	11	#上海交通大学#医学院#第九人民医院#肾内科#
VIP:34532452	1	#上海交通大学#船舶海洋与建筑工程学院#工程力学系#
VIP:34542347	1	#上海交通大学#电子信息与电气工程学院#
VIP:34693216	1	#上海交通大学#医学院#瑞金医院#内分泌代谢病科#上海市内分泌代谢病临床医学中心#
VIP:35228023	1	#上海交通大学#国际与公共事务学院#

4.3　中文机构多层级词表编制

4.3.1　基于共现矩阵与等值匹配的分块算法

利用多层级机构的三维共现矩阵,计算一级机构、二级机构和三级机构的共现频次。以船舶海洋与建筑工程学院为例,可设置共现频次阈值为3,多层级机构共现关系部分计算结果见表5。利用等值匹配的分块算法进行数据分块,将二级机构作为分块键时,在键值为“船舶海洋与建筑工程学院”的数据分块中,可通过相似度计算等发现同一机构实体的多种变异名称,如“土木工程系”“土木系”“建筑学系”“建筑系”等;将三级机构作为分块键时,在键值为“工程力学系”的数据分块中,可发现同一机构实体的简称和全称等多种变异名称,如“船建学院”“船舶海洋与建筑工程学院”。

4.3.2　基于编辑距离的相似度算法

在数据分块基础上,采用滑动窗口方法,滑动窗口宽度设为30,步长设为1。利用 Levenshtein 距离、Jaro

表 5　多层级机构共现关系计算结果(部分)

一级机构	二级机构	三级机构	频次
上海交通大学	船建学院	土木工程系	5
上海交通大学	船建学院	安全与防灾工程研究所	29
上海交通大学	船建学院	工程力学系	3
上海交通大学	船舶海洋与建筑工程学院	海洋工程国家重点实验室	61
上海交通大学	船舶海洋与建筑工程学院	交通研究中心	7
上海交通大学	船舶海洋与建筑工程学院	交通运输与航运系	3
上海交通大学	船舶海洋与建筑工程学院	土木工程系	40
上海交通大学	船舶海洋与建筑工程学院	土木系	4
上海交通大学	船舶海洋与建筑工程学院	安全与防灾工程研究所	4
上海交通大学	船舶海洋与建筑工程学院	工程力学系	39
上海交通大学	船舶海洋与建筑工程学院	工程管理研究所	3
上海交通大学	船舶海洋与建筑工程学院	建筑学系	11
上海交通大学	船舶海洋与建筑工程学院	建筑系	9
上海交通大学	船舶海洋与建筑工程学院	水下工程研究所	3
上海交通大学	船舶海洋与建筑工程学院	水声工程研究所	4
上海交通大学	船舶海洋与建筑工程学院	海洋岩土工程研究中心	3
上海交通大学	船舶海洋与建筑工程学院	高新船舶与深海开发装备协同创新中心	11

距离、Jaro-Winkler 距离分别测度字符串相似度,设置相似度阈值,完成机构名称归并。当前缀字符个数大于等于 1 时,Jaro-Winkler 距离在 Jaro 距离的基础上进行调整,刻画前缀相同部分的字符串相似度。基于编辑距离的字符串相似度计算结果(见表 6)。将 Jaro-Winkler 距离阈值设为 0.75 时,可获取相似度较高的字符串对,并将其作为人工验证的初始数据。最终,编制的高校机构名称多层级词表共有 3 528 个机构名称,其中,变异名称 2 834 个,规范名称 694 个。二级机构规范名称 160 个,包括学校与其他机构的共建平台 66 个;三级机构规范名称 478 个;四级机构规范名称 56 个,四级机构主要为研究所和系(如安泰经济与管理学院经济学院经济系)。

表 6　基于编辑距离的字符串相似度计算结果

序号	字符串 1	字符串 2	Levenshtein 距离	Jaro 距离	Jaro-Winkler 距离
1	土木系	土木工程系	2	0.689	0.751
2	建筑学系	建筑系	1	0.917	0.933
3	工程力学系	工程管理研究所	5	0.562	0.650
4	港口、海岸及近海工程系	港口与海岸工程系	4	0.837	0.902
5	海洋工程国家重点试验室	国家海洋工程重点实验室	5	0.839	0.839
6	国际航运系	建筑学系	4	0.483	0.483

4.4　机构名称归一

通过在上海交通大学 233 998 条数据中随机抽取 1 000条机构归一化结果数据,进行人工验证,最终得到机构名称归一化实验结果(见表 7)。经人工验证,机构名称实体边界识别错误 8 个,其中,"中心#医院#"和"医学院#医院#"等,因未进行相邻机构标识词合并处理而导致识别错误。上海交通大学一级机构未被识别的数据有 5 条,一级机构识别错误 1 个。上海交通大学二级机构识别错误 2 个,分别因为实体边界识别错误和机构多层级词表错误。实验结果表明,机构名称实体边界识别准确率为 99.2%;上海交通大学一级机构识别准确率为 99.9%,召回率为 99.3%,F 测度为 99.6%;二级机构识别准确率为 99.7%,召回率为 95.5%,F 测度为 97.6%。在二级机构未被识别的 31 条数据中,6 条数据因其变异名称未列入机构多层级词表而导致未被识别,如"上第九人民医院"和"江苏省苏州市九龙医院";5 条数据因一级机构未被识别而导致二级机构未被识别;其余 20 条数据仅署名上海交通大学缺失二级以下机构相关信息,由于仅利用题录数据中作者机构字段,因而无法对这部分数据进行二级机构识别,需要借助作者信息等进行识别。

通过在常熟理工学院 11 569 条数据中随机抽取 1 000条机构归一化结果数据，进行人工验证，实验结果表明，一级机构识别准确率为 100%，召回率为 100%，F 测度为 100%；二级机构识别准确率为 99.8%，召回率为 80.5%，F 测度为 89.1%。在二级机构未被识别的 148 条数据中，全部为仅署名常熟理工学院。

表 7　机构名称归一化实验结果统计

命名类型	机构名称	机构层级	机构识别正确个数(TP)	机构识别错误个数(FP)	机构未识别个数(FN)
大学	上海交通大学	一级机构	689	1	5
大学	上海交通大学	二级机构	662	2	31
学院	常熟理工学院	一级机构	761	0	0
学院	常熟理工学院	二级机构	612	1	148

5　结论

大数据时代，学术大数据的新特征召唤机构识别模式创新，从语言学视角审视机构识别，从同一关系、相继关系和隶属关系三大识别维度出发，综合考虑模型的输入、过程和输出三个层面，将基于规则、统计的自动识别和人工验证两种方式相结合，以改善大数据环境下科研管理、学科评价、学科服务中的数据可靠性问题为导向，构建数据驱动的机构名称归一化模型，重构传统的基于数据清洗平台的自动指派＋人工指派两阶段模型，是新时代推进机构识别的科学策略。

所构建的机构名称归一化模型主要从输入数据、归一化过程和机构词表编制上进行了探索，具体体现在以下几个方面：

一是输入数据的质量控制。所构建的模型以作者机构字段为唯一数据来源，辅助规则和具有权威性和准确性的机构代码表等知识，保障了模型输入端的数据可靠性，避免了参考作者、邮编[6]等信息时因作者歧义、邮编不规范等造成的数据污染。

二是白箱模型。所构建的模型采用精确匹配法识别机构，克服了机器学习算法的黑箱局限和难以进行修正的不足。精确匹配法的核心和基础是编制的中文机构多层级词表，而词表制作是自动识别和人工验证相结合，在一个数据集中变异名称数量是有限的且能够全部被人工验证，如此产生的词表代表着学科馆员等的识别水平，并保证词表的准确性。尤其是人工验证或实际应用中发现识别错误样本时，可通过修改词表高效地完成批量纠错。

三是基于实体关系识别的机构多层级词表编制。提出基于共现关系与等值匹配的分块算法，采用基于编辑距离的相似度算法，自动识别机构实体间的隶属关系和同一关系，减少了机构多层级词表制作的人工成本，也避免了基于关键词词频统计的方法存在的多层级机构间的相互干扰的不足。针对机构实体间的相继关系等的识别和更新，可进行机构多层级词表的常态加工维护，并依据机构变革的报道信息、定期发布的内部机构代码表、文献题录数据提取等多种途径进行及时更新，降低了仅基于文献题录数据提取机构相继关系的时间滞后影响。

实证结果表明，机构名称归一化模型可实现大学命名和学院命名两种类型机构名称实体边界识别和二级机构识别，从而验证了模型的有效性，对其他类型机构名称归一化有一定的启发意义。同时，一些问题有待进一步解决，如将模型输出设为文献归属到三级以下机构时的实证验证，相继关系的自动识别，在其他文献题录数据集上验证模型的有效性等。

参考文献：

[1] 贾君枝，曾建勋，李捷佳，等. 科研机构名称归一化实现[J]. 图书情报工作，2018，62(13)：103－110.

[2] 曾建勋，王立学. 面向知识评价的规范文档建设方法[J]. 图书情报工作，2012，56(10)：101－106.

[3] 曾建勋，贾君枝. 机构名称规范数据的语义模型构建[J]. 大学图书馆学报，2019，37(1)：42－47.

[4] 刘兵. 情感分析：挖掘观点、情感和情绪[M]. 刘康，赵军译. 北京：机械工业出版社. 2017(7)：1.

[5] 沈嘉懿，李芳，徐飞玉，等. 中文组织机构名称与简称的识别[J]. 中文信息学报，2007(6)：17－21.

[6] 杨波，杨军威，阎素兰. 基于规则的机构名规范化研究[J]. 现代图书情报技术，2015(6)：57－63.

[7] 胡万亭，杨燕，尹红风，等. 一种基于词频统计的组织机构名识别方法[J]. 计算机应用研究，2013，30(7)：2014－2016.

[8] 买合木提・买买提，王路路，吐尔根・依布拉音，等. 基于条件随机场的维吾尔文机构名识别[J]. 计算机工程与设计，2019，40(1)：273－278.

[9] 杨瑞仙，毛一雷. 面向知识评价的我国科研机构命名识别方法研究[J]. 情报杂志，2015，34(7)：179－183.

[10] 杨奕虹，李雅萍，张立丽，等. 机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用[J]. 数字图书馆论坛，2013(6)：57－63.

[11] 孙海霞，李军莲，吴英杰. 基于 K-means 的机构归一化研究[J]. 医学信息学杂志，2013，34(7)：41－44＋71.

[12] 申德荣，寇月，聂铁铮，等. 实体识别技术[M]. 北京：机械工业出版社. 2017(9)：45－50.

[13] JARO M A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida [J]. Journal of the A-

merican statistical association, 1989,84(406):414 -420.
[14] WINKLER W E. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage [C] //Proceedings of the section on survey research methods, Washington,DC:American statistical association,1990:354 -359.

作者贡献说明：
杨昭：负责论文选题，研究框架设计，论文写作；
任娟：负责资料收集与整理，论文写作。

Research on Institution Name Normalization Based on Chinese Bibliographic Data

Yang Zhao[1] Ren Juan[2,3]
[1] Shanghai Jiao Tong University Library, Shanghai 200240
[2] Shanghai Publishing and Printing College, Shanghai 200093
[3] Shanghai Research Institute of Publishing and Media, Shanghai 200093

Abstract: [**Purpose/significance**] In the era of big data, institution name data presents new features such as mass, dynamic and diversity. Normalization of institution name can improve the reliability of data in scientific research management, subject evaluation and subject service under big data environment, and improve the quality and application effect of data retrieval based on institution name. [**Method/process**] From the perspective of linguistics and model construction, this paper studied name normalization. This paper constructs a Framework Model for Normalization of Institutional Names Based on Co-occurrence Relations and Similarity. Firstly, it proposed a method of identifying the entity boundary of names. Secondly, it compiled a multi-level vocabulary and proposes a normalized method of names. Finally, the Chinese bibliographic data from 2008 to 2018 were selected for experiment. [**Result/conclusion**] Experiments verify the validity of the model, which has some enlightening significance for the normalization of the names of other types of institutions.

Keywords: institution name normalization model construction big data entity boundary recognition